DOCUMENT RESUME

ED 279 694 TM 870 081

AUTHOR Resnick, Daniel P.; Resnick, Lauren B.
TITLE Understanding Achievement and Acting To Produce It: Some Recommendations for NAEP.
PUB DATE 19 Jan 87
NOTE 10p.; One of 46 papers commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report "The Nation's Report Card" (TM 870 049). For other papers in this group, see TM 870 050-094.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Achievement Tests; Aptitude Treatment Interaction; *Data Collection; Educational Assessment; Elementary Secondary Education; Evaluation Utilization; *Information Needs; Measurement Objectives; *National Surveys; *Predictor Variables; School Effectiveness; State Surveys; Testing Problems; Testing Programs; *Test Use
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

Recent developments in the National Assessment of Educational Progress (NAEP) have addressed criticisms of previous assessments and have focused on state education departments as the major clients for test results. NAEP currently indicates progress in academic learning but does not define which aspects of schools are responsible for successes and failures. Recent reports indicate that NAEP plans to assess the association of input variables and mediating variables and school performance, in order to make the program more useful to policymakers. Sophisticated methods such as path analysis and structural modeling will permit many variables to be analyzed simultaneously. The variables to be studied should be actionable or practical to carry out, should reflect a reasonable theory about what matters in producing achievement, and should specify a time scale of expected effects. Actionable variables include teacher educational background and salary, instructional time, textbooks, class size, classroom homogeneity, homework, tracking, and courses taken. Mediating and interacting variables include attitudes, social background, and academic ability. Additional attention to data analysis is suggested as the most important of the recommendations. The appendix summarizes criticisms of NAEP and its current plans. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED279694

Understanding Achievement and Acting to Produce it:

Some Recommendations for NAEP

Daniel P. Resnick and Lauren B. Resnick

Carnegie-Mellon University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. P. Resnick

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☑ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper commissioned by

THE STUDY GROUP ON THE NATIONAL ASSESSMENT OF STUDENT ACHIEVEMENT

1986

TM 870 081

BEST COPY AVAILABLE

Understanding Achievement and Acting to Produce It:
Some Recommendations for NAEP

Daniel P. Resnick, Carnegie Mellon University
Lauren B. Resnick, University of Pittsburgh

Several of the reports we have written presented assessment as an important instrument for raising standards in the schools. Standards, in our view, referred to both expectations for performance and real levels of achievement in school-based fields of study. To raise standards meant, we argued, to increase the exposure of students to well-conceived curriculum and to raise the demand for success in the mastery of that material for as many students as possible. (Resnick and Resnick, 1983, 1985). Assessment of student achievement, however, has many forms for its many audiences. When we addressed syllabus-based examinations as a form of assessment, we were thinking of teachers, principals and students as the major interested clients.

If we turn to another kind of assessment, state-wide minimum competency testing, the major client is likely to be the state legislature or department of education that mandated it. Because there is generally no curriculum tied to state-wide minimum competency testing, this kind of assessment cannot focus on the mastery of a classroom-based syllabus. Nonetheless, superintendents, principals, teachers and parents pay attention to the results because of the publicity given to the scores, the political context in which such testing is pursued, and because entire classes of each school form the test population. No school can afford to ignore the demands of such testing, even when they are only for minimum levels of achievement, on small parts of a school curriculum, sampled by multiple-choice questions. The publicity given to the scores, the implicit ranking of schools, the funding implications of poor performance, and the effects on individual students will be sure to draw attention to the results.

The National Assessment of Educational Progress (NAEP) offered a third way to measure the effect of schools on the learning of our young people. In its first twenty years, 1963-83, NAEP sampled with its own tests specific subject area performance of 9-, 13- and 17-year olds. Results were aggregated over classrooms, districts and states to produce a report card for the nation. This report card was addressed to all those - parents, educators, journalists, politicians, industrialists, opinion-makers - who might have an interest in what could be shown about the knowledge and skills of American youth.

How was it to be useful, to whom? Some of the criticisms of NAEP during these twenty years dealt the with the design and validity of the testing instruments; others with the sampling pattern; still others with the way in which the data was reported. The recurring theme of this criticism was that the results were not useful to educational policy-making. Despite its boldness in seeking to measure learning across an entire nation, the old NAEP had not targeted a specific set of educational policy-makers as its primary audience, and it had not presented data to them in usable form.

The new NAEP (1983 -) attempted to address most of these criticisms [See Appendix]. With enormous psychometric talent, the new NAEP at ETS, now in its third year of operation, made changes in test design, sampling, and reporting procedures. Behind these changes was a clearer conception of the primary users of the data collected. State education departments are now conceived as a major client for test results. Information aggregated at a state level will be shared with state education departments. In addition, complementary services have been developed which will allow state officers to undertake

evaluation of achievement within states in greater depth, using NAEP items, and with reference to NAEP norms(Messick, Beaton and Lord, 1983; Anrig and Lapointe, 1983; Pipho, 1984). We consider below some ways in which NAEP's service to states - and concomitantly to school district policy-makers - might be enhanced.

Making the new NAEP more useful to policy-makers

NAEP presently provides indicators of how well American schools are doing in producing academic learning. It does not tell us what aspects of the way schools are run are responsible for successes and failures in achievement. Thus, simply making comparative scores available on a state-by-state (or even district -by-district) basis, will not tell educators or legislators what to do to produce improvements. It may tell them to "try harder", but not what kinds of tries are likely to make a difference. With the addition of information on certain educational inputs and mediating variables, NAEP could become a tool for understanding how policies on matters such as certification, salary, class size, scheduling and textbook selection affect student achievement.

Recent reports from ETS on plans for the new NAEP suggest that plans are already underway for collection of additional background data, and much of what we have to suggest may in fact be included in those plans --which were, however, not fully specified in the documents to which we had access. Our comments should not be understood as a criticism of current plans. Rather we intend to describe here the criteria that we think should be applied in designing and carrying out this expansion of NAEP's activity.

Beyond knowing <u>whether</u> American schools are succeeding in producing the levels of achievement hoped for, policy-makers need to know what factors affect achievement levels. The natural variations in achievement between classrooms and schools provide a kind of laboratory for developing this knowledge. If data on input and mediating factors are collected appropriately, analyses can plot the extent to which variations in input variables are associated with variations in output. Sophisticated methods (path analysis, structural modeling) permit many variables to be analyzed simultaneously, so that their interactions as well as their independent effects on achievement can be assessed and so that reasonably strong inferences about causal patterns can be made. Such analyses will reveal some input variables to be much more regularly associated with achievement than others. These would become candidates for educational policy action.

Because NAEP is based on light sampling, it cannot (without huge increases in cost) assess actual classroom processes. Nor is NAEP well suited to detailed diagnosis of the reasons that individual students have trouble with particular types of learning. For these reasons, it is hard to imagine using it for decisions on how to tailor instruction at the school, classroom or individual student level. NAEP can, however, be very useful for making decisions that are somewhat more distant from the act of instruction. For this reason, the proper audience for NAEP input studies is state and district level officials.

What input variables should be studied?

The range of variables that might be included in NAEP data collection is enormous, and some principles are needed to select among the many candidates. We suggest three such principles here.

First,the variables of primary interest must be actionable by state and school district level actors. A great deal of research on educational effects focuses on variables that educators and education policy-makers can do nothing about. We learn repeatedly, for example, that parents' education and income heavily determine who will be high achievers in school. But these are not actionable variables; educators cannot do anything to change students' socio-economic status. As a result, such information alone cannot effectively guide educational policy or practice. NAEP data collection and analyses should focus on variables that educators can do something about.

Secondly,the variables chosen should reflect a reasonable theory or hypothesis about what matters in producing achievement. Structural models and path analyses do not work well as "hypothesis-generating" techniques. There must be one or several hypotheses to be tested for the methods to be applied successfully. Further, NAEP will necessarily have limited resources for collecting measures of input variables and conducting analyses. For both of these reasons, variables included in the input studies should be limited in number and chosen on the basis of persuasive hypotheses about probable educational effects. Multiple hypotheses can be tested concerning which variables are key, and not all variables must be included in every analysis, so this proposal does not restrict NAEP to a single hypothesis. But it does call for reasoned theoretical analysis in advance rather than a raw empiricist ("collect everything and then we'll see") approach.

Thirdly, there must be specification of the time-scale of expected effects. Some of the variables that can be measured are in areas where changes in policy can bring about measurable short-term effects --such that, for example, changes in achievement could be noted by the next cycle of NAEP testing. Others may have only long-term effects (changing teacher salaries, for example, would likely affect achievement only slowly, through long-term effects on who is recruited into and retained in the teaching profession). Both short-and long-term effects are of interest, but it is important that in planning NAEP input studies, planners be clear about the time-scale of effects they are predicting or expecting.

Some Variables That Are Actionable

We discuss here examples of the kinds of variables that would meet these criteria.

- Educational background of teachers. For example, highest degree, type of institution, major field of study, year of entry into the profession, association memberships, courses taken in the subject matter taught. These are the kind of elements usually specified in certification requirements. Certification requirements are actionable. Changes in them will in the long term produce changes in the qualifications of the teaching force. To make appropriate changes in requirements, policy-makers need to know such things as whether achievement is more affected by level of training in one's teaching discipline or by the extent of training in education methods (to pick a currently debated issue). Data on teacher education, properly analyzed, can yield such information.
- Salary. This is clearly something over which policy-makers have control. It seems implausible that higher salaries would be able to directly improve students' achievement. But indirectly salaries might have a large effect by--over time--changing the composition of the teaching force. One would probably want to test theories in which salary is thought to affect

educational background of teachers, which in turn affects student achievement.

- Instructional time variables. For example. length of school day, length of school year, and, most important, number of hours per year of instruction in a particular subject matter. These are all measures of what researchers call allocated time. We know that what really matters in learning is engaged time--the time that students actually attend to the kinds of activities that produce learning. We choose measures of allocated time for NAEP studies, however, because state and local level policy-makers are unlikely to be able to act directly upon engaged time; and because the conditions under which NAEP data is collected make it unlikely that good estimates of engaged time can be made by NAEP.
- Textbook used. This is actually a stand-in for curriculum. It seems to be the best indicator NAEP is likely to be able to identify to assess the curriculum to which children are actually exposed. If more sensitive indicators of curriculum could be made feasible, we would prefer using them. In any case, what we are looking for is some estimate of the content of the curriculum that may produce certain achievement effects.
- Class size. This is a variable over which policy-makers have considerable control and which is widely believed to affect achievement. Including it as an input variable might shed some light on whether and when class size matters. There is a good chance that class size actually matters more for weaker students and for students from less educationally-favored families. For this reason, one would want to examine the effect of class size along with an ability assessment (see below).
- Homogeneity/heterogeneity of classes. This, too, is a variable over which policy-makers have some control and which is widely believed to affect achievement.
- Homework. This is a variable over which policy-makers have some control and which is widely believed to affect learning. It would be useful to refine our knowledge of its effects. What kinds of homework, in what subject fields, for example, are likely to have the greatest effect? To learn this, teachers will have to be asked to describe the nature and amount of homework they give, how they grade it, and the like.
- Tracking. This is clearly a variable over which policy-makers have control. Tracking influences the kinds of curriculum to which students are exposed and the amount of time they can devote to different topics. The effects, however, may be different and more dramatic for the very brightest students and the weakest than they are for those in the middle. For this reason, it would be important to examine tracking effects along with an SES assessment.
- Courses taken. Courses taken are influenced by tracking decisions, but not completely controlled by them. Therefore, a direct report of courses actually taken should be made. Analyses of this data can inform action on state and district course requirements.

Mediating and Interacting Variables

Mediating variables. It is likely that many of the input variables over which policy-makers have control have their effects --if any--by the way in which they influence certain mediating variables. For example, class size may not act directly to improve achievement, but it may affect children's and teachers' attitudes toward one another, and thereby indirectly affect achievement. NAEP could collect data on a limited number of such mediating variables in order to enrich the input-achievement analyses. Promising mediating variables include students' stated interest in the course or subject matter, students' expected grade in the course, students' attitude toward the teacher, teacher's expectation for student performance, and the like.

Interacting variables. Many of the input variables may affect different kinds of students or schools differently. The most obvious cases are those where students of different social background or academic ability respond differently to different texts, to class size, etc. A considerable research literature points to

such interactions, and it would be wise to include measures of the most likely interactive variables in the NAEP studies in order to make the input studies as useful as possible. NAEP should look for very simple measures that are adequate "stand-ins" for more precise measures of the potential interacting variables. For example, aptitude for school learning could be assessed by asking students to record the grades they received the previous year.

Some caveats

There are some limits and complications to what NAEP can be expected to reveal from input studies of the kind we are suggesting. The important limits to be kept in mind are that NAEP tests only periodically, and that it samples lightly. Together these two conditions mean that it will not be possible to tie achievement to instructional input variables as tightly and elegantly as might be wished. Consider, for example, the difficulties deriving from periodic assessment. Students are tested, except for reading, only every four years, and each assessment is meant to measure the cumulative effects of at least the preceding four years of instruction. Which of the past four years of instructional variables (texts, homework, allocated time, teachers, etc.) should be assessed as input? Ideally, all--but practically, only the current year has much chance of being reliably reported. For this reason, we sugggested limiting input variable reports to the current instructional year. However, it must be recognized that this will provide only a general indication of the kind of instruction that has been experienced, rather than a precise measure.

The four-year testing cycle poses particular problems in the 11th grade assessment. In science, for example, the test measures a cumulation of skill and knowledge gathered across a number of different specialized courses (biology, chemistry, etc.). But reports on input variables will be for only one of these courses. One must expect lower correlations between instructional variables and achievement than would be found if the achievement items were limited to the course on which inputs are reported. Despite these limits, we believe that NAEP can provide useful information about the kinds of variables that are actionable by policy-makers and that affect achievement.

Finally, and most important, our recommendation is not just that more data be collected by NAEP, but that more attention than in the past be paid to data analysis. Even if all our suggestions for extended collection of data on input variables were followed, the effort would be of little value without the very sophisticated analyses that we are assuming will be carried out. Efforts of this kind call for the best statistical and interpretive talent the nation has to offer. Such talent should be sought not only within ETS, but in other research groups that may cooperate in the broad NAEP undertaking. This will not be cheap. But without it, we will learn little from NAEP about how to improve achievement.

APPENDIX

Plans for the new NAEP address quite directly the criticisms made of its predecessor.

Criticisms of NAEP (1963-83)

- Sampling procedure did not allow for the collection of representative scores for individual schools.
- State-level results were collected, but not reported. Reporting policy was to avoid publication of data that would lead to state-level comparisons. Published results were aggregated at a very high level (four regions for the United States).
- The ten subject areas tested were not always taught in school to the youngsters tested. Civics, for example, had not been taught to all the 13-year olds tested.
- Testing was of age groups rather than grades. 9-, 13-, and 17-year olds had been tested instead of 3rd, 7th and 11th graders. The advantage was that stragglers who would not have been promoted to those grades were included. In the real world, however, school grouping was by grade rather than age.
- There was no clear definition of standards. Previous test results in subject areas remained the norm by which current year results were measured. Results were reported in terms of percentage right on individual items. It was difficult to translate this kind of reporting into useful knowledge for educational decision-making;
- There was an unanswerable doubt about the construct validity of the testing instrument for the fields under examination, because no effort was made to relate peformance on NAEP items to peformance on other subject area measures of achievement.
- Plans to sample the performance of age cohorts were inadequate; it was not possible to follow the progress over time of specific birth cohorts in specific subject areas.

Plans for the New NAEP (1983-)

- Testing of different domains of knowledge will be pursued with more attention to the construct validity of the testing instrument. Test items will be compared for performance with other measures of achievement in similar subject domains.
- Standards for achievement will be established by a consensus of experts for each subject area. Results from a previous NAEP testing will no longer be the only benchmark for standards.
- Light matrix sampling will yield more satisfying results because of new techniques (IRT Scaling, BIB) that allow performance on NAEP items or clusters of items to predict performance on both non-NAEP measures and other NAEP batteries (Lord, 1980; Messick, 1985).
- Although some age-group testing will be maintained for equating purposes and to maintain the value of earlier series data, testing will be directed to specific grades rather than age groups. The grades tested will be third, ninth and eleventh.
- Specific age cohorts will be tested at four-year intervals in specific subject areas to provide information about the learning of those born in the same time period.
- Background information about schools and pupils will be gathered on a broader range of questions and more systematically by questionnaire and field-staff inquiry. The focus of the inquiry, for students, will be values, use of time, and socio- economic data. From the schools, information will be collected on issues like racial mix, tracking patterns, school size, and community characteristics. The goal of this data collection is explanation of the factors responsible for school success or failure, using path analysis.

- Four central areas of the curriculum will be tested at two- and four-year intervals, with more systematic attention to school-taught subjects. Reading will be tested every two years; writing, math and science every four years, in alternate cycles, beginning in 1985-86.
- The primary sampling units remain "typically counties", with the aggregation of smaller county units so that there are at least 1500 students of each age group in each sampling unit. Only modest changes in sampling patterns were anticipated to accomodate new psychometric techniques (i.e., BIB spiraling). Modest representation of individual schools will continue, with typically 16 to 25 students at each age/grade for each subject tested in a participating school (Messick, Beaton, Lord, 1983, pp. 22-23).

REFERENCES

Anrig, G. and A. Lapointe."NAEP: The Nation's Report Card." *Phi Delta Kappan* 65(1983), 52-54.

Messick,S. "Response to Changing Assessment Needs: Redesign of the National Assessment of Educational Progress." *American Journal of Education* 94(1985), 91-105.

Messick, S., A. Beaton and F. Lord. *National Assessment of Educational Progress Reconsidered: A New Design for a New Era.* NAEP Report no. 83-1. Princeton, N.J.: National Assessment of Educational Progress, 1983.

Piphc, Chris, "Stateline: Getting Schools on the Scoreboard." *Phi Delta Kappan* 66(1984), 237-38.

Resnick, D. and L. Resnick. "Improving Standards in American Schools." *Phi Delta Kappan* 65(1983), 178-80.

Resnick, D. and L. Resnick. "Standards, Curriculum, and Peformance: A Historical and Comparative Perspective." *Educational Researcher* April, 1985, 5-20.